Atigeo at TREC 2014 Clinical Decision Support Task

Yishul Wei*, ChenChieh Hsu*, Alex Thomas, Joseph F. McCarthy
Atigeo, LLC
800 Bellevue Way NE, Suite 600
Bellevue, WA 98004 USA
joe.mccarthy@atigeo.com

Abstract

The TREC 2014 Clinical Decision Support Track task involves retrieval and ranking of medical journal articles with respect to their relevance to prescribing tests, diagnosing or treating a patient represented in a short case report. The Atigeo xPatterns™ platform supports a variety of *ensemble* methods for developing and tuning information retrieval (IR) system components for a task and/or domain using labeled data. For TREC 2014, we combine results from an ensemble of search engines, each with a configurable suite of natural language processing (NLP) components, to compute a relevance score for each article and topic. We describe our ensemble approach, the strategies and tools we use to create labeled data to support this approach, the components in our IR / NLP pipeline, and our results on the TREC 2014 CDS task.

1 Introduction

The TREC 2014 Clinical Decision Support (CDS) track was designed to assess the ability of search engines to find biomedical journal articles relevant to clinical questions about a patient. Each topic within this track consists of a sentence-long summary and a paragraph-long description of a patient case, along with one of three types of clinical information need: `diagnosis`, `test` or `treatment`. The CDS task is to retrieve a ranked set of up to 1000 documents that are relevant to a particular case – based on the summary, description, or both – which are likely to support a physician's decision on appropriate patient care, including proper diagnosis, the tests the patient should undergo, and how the patient should be treated.

The corpus for the retrieval task is a snapshot of the PubMed Central (PMC) Open Access Subset[1] on January 21, 2014. This set contains the abstracts, full texts, and other metadata of 733,138 articles in the biomedical domain, available in XML format conforming to the National Library of Medicine Journal Publishing DTD[2]. CDS participants were invited to submit up to 5 sets of ranked documents deemed relevant to 30 topics (summary, description pairs), 10 for each of the three types of information need. The 30 evaluation topics were to be considered "blind", with one sample topic summary along with three examples of relevant documents on the CDS track website[3] available for use in development and testing. The large size of the corpus and the sparseness of the development set of topics posed considerable challenges for the CDS track this year.

The Atigeo team developed an end-to-end document retrieval pipeline for the TREC 2014 CDS task and produced a set of unofficial topics together with relevance judgments for internal evaluation. The pipeline utilizes two open-source search engines – Solr/Lucene[4] and Indri/Lemur[5] – and includes several text processing and natural language processing (NLP) modules, such as negation tagging, age grouping, and semantic-based query expansion, as well as a final ensemble algorithm that combines different ranked lists to improve retrieval results.

We conducted numerous experiments with different configurations of components to determine the five runs we submitted as our official results.

---

* Work reported here undertaken while the first two authors were graduate student interns at Atigeo. Current contact information for first 3 authors: yishuwei@uw.edu, cjhsu@uw.edu, alex.thomas@atigeo.com

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] http://dtd.nlm.nih.gov/publishing/
[3] http://www.trec-cds.org/
[4] http://lucene.apache.org/solr/
[5] http://www.lemurproject.org/indri.php

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Atigeo at TREC 2014 Clinical Decision Support Task	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Atigeo LLC,800 Bellevue Way NE, Suite 600,Bellevue,WA,98004	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

The TREC 2014 Clinical Decision Support Track task involves retrieval and ranking of medical journal articles with respect to their relevance to prescribing tests, diagnosing or treating a patient represented in a short case report. The Atigeo xPatterns??? platform supports a variety of ensemble methods for developing and tuning information retrieval (IR) system components for a task and/or domain using labeled data. For TREC 2014, we combine results from an ensemble of search engines each with a configurable suite of natural language processing (NLP) components to compute a relevance score for each article and topic. We describe our ensemble approach, the strategies and tools we use to create labeled data to support this approach the components in our IR / NLP pipeline, and our results on the TREC 2014 CDS task.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

The following sections provide more details about our approach.

2 Methodology

In Sections 2.1 and 2.2 we describe the indexing and retrieval strategies for the Indri-based search pipeline of our system. The indexing strategy for the Solr-based pipeline is similar. Our query expansion strategy, however, relies heavily on the Indri query language, which is not compatible with Solr. Section 2.3 describes the ensemble algorithm that combines the search results from Indri and Solr to produce the final results.

2.1 Indexing Strategy

For each document in the corpus, we extract its title, abstract, keywords and body, and index the documents by Krovetz-stemmed (Krovetz 1993) terms in these fields except for those on the stopword list developed by Atigeo for its clinical autocoding (CAC) product.

We constructed several indexes wherein documents were preprocessed to different extents. We utilize three independent preprocessing modules: negation tagging, string normalization, and age grouping. The negation tagging module involves identifying negated terms using the open-source program NegEx[6] and prepending an "nx" prefix to the negated terms, following Limsopatham (2011). The string normalization module removes non-alphanumeric characters. Acronyms with periods and hyphenated words thus become single terms. (The default setting of Indri is to break them into several terms.)

The age grouper first matches age-description phrases (e.g. "a 30-year-old woman") with a few hand-designed rules, and then replaces them with an age-group identifier. We divided ages into groups of 0–10 years old, 10–20 years old, 20–30 years old, and so on. The rationale for age grouping is that a document with a case description should be regarded as relevant if the patients described in the document and the query are in the same age group and share similar clinical conditions, even though their exact ages are different. Without age grouping, a 34-year-old appears as dissimilar to a 36-year-old as to a 70-year-old to the search engine, since in either case the age-relevant string does not match. In contrast, with age grouping, the 34-year-old and the 36-year-old would be regarded as more similar since they are in the same age group.

Since the three preprocessing modules can be independently turned on or off, we created a total number of 8 indexes including the basic index (where documents were indexed without any preprocessing) that we could conduct our experiments on.

2.2 Retrieval Strategy

At the retrieval stage, we experimented with using either topic summaries or descriptions as the queries submitted to the search engine, but not both, since we assumed that all information in the summary would also be contained in the description (although with no sample descriptions, we were unable to validate this assumption during development). The queries are preprocessed with the same three modules as described above, depending on the index being used. For example, when querying the index with tagged negated terms, we also apply the negation tagger to the queries.

Query expansion has been shown to be an effective strategy for improving search results in biomedical information retrieval (Srinivasan 1996a, 1996b; Aronson & Rindflesch 1997). Recognizing the fact that clinical descriptions in the queries might use different terms than those used in the documents when referring to similar concepts, we developed a semantic-based query expansion module in our search pipeline. The idea is to add to the queries the synonyms of the clinical terms present in the query. To achieve this goal we used MetaMap (Aronson 2001), a widely known tool for extracting clinical concepts from free text. For each query, MetaMap maps the clinical terms it identifies to Medical Subject Headings (MeSH), a comprehensive controlled vocabulary representing many concepts described in biomedical journal articles.

The query expansion module then queries the Unified Medical Language System (UMLS) database to get all the string variants of the concepts. These string variants are then added to the queries, using the phrase-level matching and synonym operators of the Indri query language. Specifically, each string variant that comprises more than one word is wrapped inside a phrase-level matching

[6] http://code.google.com/p/negex/

operator `#1(...)`, and all string variants of a single concept are grouped as synonyms using the synonym operator `{...}`. The reason for crafting the expanded queries this way is that if we simply added the string variants as independent terms, we would inflate the weight of the concepts with more string variants. The synonym operator enables us to specify the variants to be matched in the documents while maintaining the total weight for each concept to be the same as a single term.

The following example shows how a summary from the official topics (Topic 13) is transformed by different modules of our system. The original summary is:

```
30-year-old woman who is 3 weeks
post-partum, presents with short-
ness of breath, tachypnea, and
hypoxia.
```

After preprocessing (with string normalization and age grouping; there is no negated term in this summary) and removing stop-words, we have:

```
threenxage woman who is weeks
postpartum presents shortness
breath tachypnea hypoxia
```

Running MetaMap on the original summary returns the following concepts:

```
C0043210: Female (Woman) [Popula-
tion Group]
C1148523: Parturition (Child-
birth) [Organism Function]
C0013404: Shortness of Breath
(Dyspnea) [Sign or Symptom]
C0231835: Tachypnea [Sign or
Symptom]
C0242184: Hypoxia [Pathologic
Function]
```

After querying the UMLS database, the string variants of all concepts are combined with the preprocessed summary to form the final query:

```
#combine(threenxage woman who is
weeks postpartum presents short-
ness breath tachypnea hypoxia
{ woman women }
{ childbirth
  #1(human parturition function)
  parturition }
{ dyspnoea
  #1(shortness of breath)
  breathlessness
  dib
  #1(breathlessness nos)
  sob
  #1(difficulty breathing)
  breathless
  #1(respiration difficult)
  dyspnea }
{ #1(rapid respiration)
  tachypnea
  tachypneic
  tachypnoea
  #1(rapid breathing) }
{ hypoxic
  #1(oxygen deficiency)
  hypoxia
  #1(decreased oxygen supply) } )
```

This query is then sent to the Indri search engine to query the indexes of the properly preprocessed documents.

2.3 Ensemble Re-Ranking

Indri and Solr both implement state-of-the-art information retrieval algorithms that deliver high-quality search results. To further improve them, we developed an ensemble framework to combine different ranked retrieval results into a single ranked list. We believe each information retrieval model encapsulated by Indri and Solr has its own advantages. The logic of the ensemble is thus to keep all the documents that are predicted to be highly relevant by either search engine.

As an illustrative example, suppose the two search engines retrieve four documents: A, B, C and D. Document A is ranked as highly relevant by both Indri and Solr, Document B is ranked as highly relevant only by Indri, Document C is ranked as highly relevant only by Solr, and Document D is ranked as only marginally relevant by both Solr and Indri. We believe the optimal aggregate ranking for these four documents should be "A > B ≥ C > D" or "A > C ≥ B > D". To model this idea, our system uses the following formula below to calculate the final score of each document:

$$\frac{SolrWeight}{SolrRank} + \frac{IndriWeight}{IndriRank} + \frac{1}{SolrRank \times IndriRank}$$

where *SolrWeight* and *IndriWeight* are parameters that represent our relative confidence of each

search engine. For most common cases, the sum of the two weights should be 1.

3 Unofficially Labeling Unofficial Topics

One of the challenges we faced in participating in the inaugural offering of the CDS track in TREC 2014 was the lack of labeled data. A single sample `diagnostic` topic summary and three sample relevant documents were provided on the TREC 2014 CDS homepage. No sample summaries for `test` or `treatment` topics were provided, nor were any sample descriptions of any topic types made available to participants.

A post on the TREC 2014 CDS Google Group suggested participants might find potentially useful examples of short case histories at CasesDatabase.com[7] and the topics provided for the ImageCLEF 2013 medical task[8]. We discovered that the single sample topic provided on the CDS track website was a topic from the ImageCLEF collection.

3.1 Creating Unofficial Topics

We created two sets of unofficial topic summaries for unofficial evaluation during our development efforts. We decided not to create any topic descriptions because there were no examples to guide us.

One set of topic summaries was derived from CasesDatabase.com. We randomly selected cases to review for potential relevance and generality; those that seemed sufficiently general were used as "seeds" from which we derived 3 topics, one for each type of information need. This derivation of 3 topics from a single case was done to reduce the cognitive overhead of labeling the search results: rather than having to analyze *n* cases when judging results, we could analyze *n*/3 cases, with some additional overhead due to the difference between the topic types (see below).

For a given seed case, the `test` topic we created contains patient demographics, history and signs and symptoms at the time of presentation of a chief complaint. Our interpretation of a `test` topic was that it should represent a scenario in which a physician would be searching for relevant information to help decide which tests to prescribe for a patient. It would thus not include any test results, diagnoses or treatments. The `diagnosis` topic includes all the information from the `test` topic, plus any results of any tests presented in the case, and the `treatment` topic includes everything in the `diagnosis` topic plus any diagnoses included in the case. The following topics illustrate 3 topics derived from the same case:

> **Test:** A 43-year-old Caucasian woman who suffered from chronic menorrhagia was started on triptorelin, a gonadotrophin-releasing hormone analogue. Three days later, she developed gradually worsening headaches accompanied by bilateral visual disturbance.

> **Diagnosis:** A 43-year-old Caucasian woman who suffered from chronic menorrhagia was started on triptorelin, a gonadotrophin-releasing hormone analogue. Three days later, she developed gradually worsening headaches accompanied by bilateral visual disturbance. Examination revealed bilateral papilledema and enlarged blind spots on her visual fields.

> **Treatment:** A 43-year-old Caucasian woman who suffered from chronic menorrhagia was started on triptorelin, a gonadotrophin-releasing hormone analogue. Three days later, she developed gradually worsening headaches accompanied by bilateral visual disturbance. Examination revealed bilateral papilledema and enlarged blind spots on her visual fields. A diagnosis of benign intracranial hypertension was made and confirmed on magnetic resonance imaging.

We initially identified 10 seed cases from which we derived 30 topics. However, in the process of rendering unofficial relevant judgments on results returned by our system for those topics, several of these cases presented significant challenges with respect to our ability to assess the relevance of articles. We discarded all 3 topics derived from some cases, and discarded 1 or 2 of the topics derived from others. We ended up with 12 topics that we deemed sufficiently "assessable" for use in our ongoing development process (which involved iteratively generating additional results from additional system configurations that would then need to be judged or labeled).

As might be expected, many of the ImageCLEF 2013 topics – all of which are of the `diagnosis`

[7] http://www.casesdatabase.com
[8] http://imageclef.org/2013/medical

type – are based on cases in which medical imaging information played a significant diagnostic role. However, we identified 3 cases with sufficient information not directly related to imaging to warrant inclusion in our unofficial topics. We derived 3 additional unofficial topics (one of each type) from 2 of the ImageCLEF topics, and included the other one directly (with no further derivations) in our unofficial topics. Our final set has 19 unofficial topics (6 `test`, 7 `diagnosis` and 6 `treatment` topics).

3.2 Labeling Unofficial Topics

In order to understand the effect of our system components and configurations on the quality of retrieval, we needed to obtain relevance labels. For determining the quality of simple components in isolation, unit tests and spot-checking may be sufficient, but more comprehensive measurements are needed when determining the quality of a system of multiple components on a large data set. A comprehensive measurement requires that the judgments be as consistent as possible, so that a comparison of different configurations of the system is meaningful. We limited our set of judges to 2 members of the team to maximize consistency, and monitored inter-rater agreement to make sure that there was a shared understanding of the topics and documents.

Another aspect of consistency is establishing an agreed upon definition of relevance. The topics in this task consist of both a case history as well as the type of clinical information need, so relevance must be defined in terms of both parts. There was some guidance from NIST on how relevance would be determined. In March, an introductory message with a task definition was posted to the TREC CDS mailing list. In June, another post provided further guidance with respect to how to interpret the `diagnosis` type. With this guidance in hand we could create some basic guidelines for each topic type.

Even with the basic guidelines we established, there were still many challenges in producing the internal judgments. There were no medical experts on the team, so for each topic substantial research had to be done to understand the case. In order to maintain a tenable workload, the top five results of each run were judged, to calculate NDCG@5.

4 Results on Unofficial Topics

There are 16 possible configurations for Indri, given all the possible combinations of using or not the four indexing and retrieval preprocessing modules (negation tagging, string normalization, age grouping, query expansion). There are only 8 possible configurations for Solr, since query expansion is not available for Solr. We tested 22 of the configurations on the unofficial topics we created (2 Solr indexes were not completed in time). The best performing configurations – shown in Table 1 – are ranked according to their normalized discounted cumulative gain (NDCG) scores averaged over all topics. We show only the configurations with average NDCG above 0.75.

Some interesting trends can be observed in the table. First, baseline Indri with no pre-processing (row 9) performs surprisingly well. Second, negation tagging has more positive effects for Solr than for Indri; in fact, negation appears to negatively impact the NDCG score for Indri. Third, we see that Solr with preprocessing but without query expansion can yield results comparable to those of Indri with query expansion. In general, different components appear to have different effects on the performance of each search engine. With some configurations Solr performs better, but with other configurations Indri wins.

Table 1: Configurations & Scores on Unofficial Topics

Rank	Search Engine	Negate?	Normalize?	Group Ages?	Expand Query?	NDCG
1	Indri	☐	☑	☐	☑	0.7847
2	Indri	☐	☑	☑	☑	0.7839
3	Indri	☐	☐	☐	☑	0.7821
4	Solr	☑	☑	☑	☐	0.7812
5	Indri	☑	☑	☑	☑	0.7780
6	Indri	☑	☐	☑	☑	0.7770
7	Solr	☑	☑	☐	☐	0.7769
8	Indri	☐	☐	☑	☑	0.7673
9	Indri	☐	☐	☐	☐	0.7667

We also ran our ensemble algorithm on every pair of configurations (with same or different search engines). This produced 462 sets of results in addition to our 22 "singleton" results. We rendered 529 unofficial judgments on 19 unofficial topics based on the aggregated set of the top 5 results from all 484 configurations.

Ensemble configurations tend to perform better than singleton configurations such as those shown in Table 1. Our best-performing ensemble (NDCG 0.7983) combines the Indri search engine with string normalization, age grouping, and query expansion with Indri using only negation tagging and query expansion. However, since many of the other configurations achieve NDCG scores very close to 0.7983, and since these are all based on unofficial labels for unofficial topics, we did not want to rely too heavily on small differences in the scores.

The configurations used for our final submissions to TREC were:

the "best" ensemble: row 2 in Table 1 plus another configuration not shown in the table (NDCG 0.7983)

the "full" ensemble: rows 4 and 5 in Table 1 (NDCG 0.7852)

the "full" Indri singleton configuration: row 5 in Table 1, (NDCG 0.7780).

Our unofficial topics only contain summaries, and so we did not run any experiments using descriptions, but we did want to see how our system would perform on both fields. We decided to allocate 4 of our allotted 5 runs to using our "best" and "full" ensemble configurations on both the summary and description fields, and allocated our 5th run to the "full" Indri singleton configuration.

Given our uncertainty about the reliability and statistical significance of small differences in our unofficial scores, we decided to designate the "full" ensemble configuration on the summary field as our official run (atigeo1). The configurations, target topic fields, scores on our unofficial judgments of unofficial topics, and corresponding labels are shown in Table 2.

Table 2: Configurations for 5 Submitted Runs

Configuration	*Target*	*Score*	*Label*
Full ensemble	summary	0.7852	atigeo1
Full ensemble	description	n/a	atigeo2
Best ensemble	summary	0.7983	atigeo3
Best ensemble	description	n/a	atigeo4
Full singleton	summary	0.778	atigeo5
Full singleton	description	n/a	-

5 Official Topics

The official topics were released April 30th; however participants are prohibited from viewing official TREC topics until after they have ceased system development, i.e., the topics should be treated as "blind" and thus not influence the development or tuning of any system. Once we submitted our runs, we examined the official topics to assess the accuracy of our expectations.

The `diagnosis` topics focus on determining the most likely diagnosis given a set of patient complaints and a patient history; some of them also include a set of test results, which violate one of the assumptions we made in developing our unofficial topics. The `test` topics focus on determining the next test that should be performed given a set of complaints and a patient history; some of these topics also include results of previous tests, which also violate our assumptions. The `treatment` topics appear to follow one of two general patterns: determining a curative or palliative treatment for some diagnosed illness or symptom, or determining a preventative treatment for a concerned patient. We had assumed `treatment` topics would follow only the first pattern.

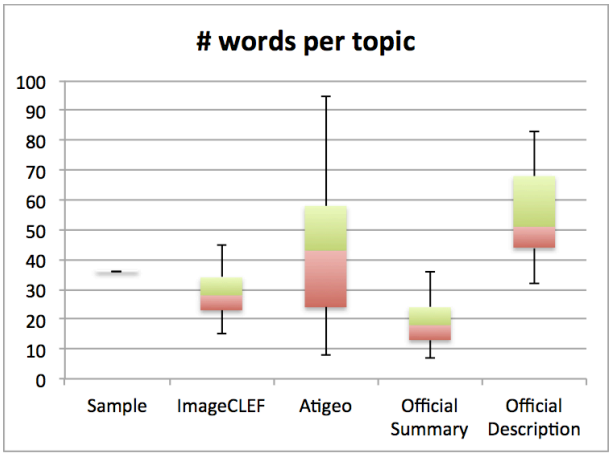

Figure 1: Lengths of Topics

We compared the lengths, in terms of tokens, of the official and our unofficial topics sets, as well as the sample topic given at the CDS track website (Figure 1). The sample topic length is not very representative of either the official summaries or descriptions, but is instead between the two. The ImageCLEF topic lengths are also between official summary and description lengths. Our unofficial

topic summaries have a wider distribution of lengths, but their median length is closer to that of official topic descriptions than official summaries.

6 Analysis of Results on Official Topics

Table 3 shows the average scores across all 30 topics for inferred Average Precision (infAP), inferred Normalized Discounted Cumulative Gain (infNDCG), R-precision (Rprec) and Precision@10 (P10) for each of our 5 submitted runs. Asterisks indicate the best performing run for each metric.

Table 3: Official Scores for 5 Runs

	atigeo1	atigeo2	atigeo3	atigeo4	atigeo5
infAP	0.0524*	0.0467	0.0513	0.0501	0.0514
infNDCG	0.1960	0.1795	0.1936	0.1891	0.1996*
Rprec	0.1601	0.1569	0.1721*	0.1618	0.1626
P10	0.3033*	0.2833	0.2967	0.2867	0.2933

Our "official" run (atigeo1) – representing the "full aggregation" configuration and targeting only the summary field – achieved the highest infAP and P10 scores (0.0524 and 0.3033) of the 5 runs, and was a very close second (0.1960 vs. 0.1996) under the infNDCG metric.

We observed a large gap between the NDCG scores for our system on the unofficial and the official topics. This difference suggests that our unofficial topics are not representative of the official topics. In examining the topics, it appears that the official summaries or descriptions tend to contain more general terms in describing symptoms or patient conditions, while our unofficial topics tend to use specific terms extracted from specific cases.

Also, as mentioned above, our unofficial topics were created in an incremental way. We assumed that `test` topics contain the least amount of information, while `treatment` topics contain the greatest amount of information. However, as we looked at the official topics, we found that this was not the case, and that the official topics usually only contain descriptions about symptoms or conditions, regardless of the query type. We believe these differences in terminology and length contribute to the substantially different NDCG scores for the two sets of topics. However, it is noteworthy that our preprocessing and query expansion modules yield consistent performance gains over the baseline search engines for both sets of topics.

7 Related Work

Many ideas explored in the submissions to the Medical Records track[9] of TREC 2011 and TREC 2012 were either similar to the approach we took in the TREC 2014 Clinical Decision Support track, or have potential to be utilized for improving the CDS system. We need to bear in mind, nevertheless, that although both were biomedical information retrieval tasks, the scopes for the Medical Records and CDS tracks are different. In TREC 2011 and TREC 2012, the Medical Records track focused on the problem of cohort selection. Given patient descriptions, that goal was to identify similar patient in a corpus of electronic medical records (EMRs). The strategies that the submissions took can be generally classified into two types, knowledge-based query formation and semantic-based query/document preprocessing, which are discussed in Sections 7.1 and 7.2, respectively. Some follow-up studies on applying more advanced NLP techniques to this problem are discussed in Section 7.3. Finally, Section 7.4 returns back to the issues in medical document retrieval in general.

7.1 Knowledge-Based Query Formation

The knowledge-based approach was first developed for clinical question answering (Demner-Fushman & Lin 2007). Under this framework, clinical queries were formulated in accordance with the guidelines of evidenced-based medicine (EBM, see Sackett et al. 1996). Specifically, each clinical query can be divided into several parts:

Clinical task, such as etiology, prognosis, diagnosis, and treatment or prevention. (This classification scheme was first proposed by Haynes et al. 1994.)

PICO elements, which stand for population/problem, intervention, comparison, and outcome (Richardson et al. 1995).

Strength of evidence, the level of confidence in the results presented in the research.

The U.S. National Library of Medicine's systems for TREC 2011 and TREC 2012 Medical Records track (Demner-Fushman et al. 2012, 2013)

[9] http://trec.nist.gov/data/medical.html

involve reformulating the cohort selection problem into a query of this knowledge-based model. An EBM-like query frame is created for each topic, which was then submitted to search engines to retrieve documents.

Although this approach achieved the best performance among all submissions to the Medical Records track, it was not an automated retrieval system. The transformation of the topics into the query frame was done by hand. In the clinical decision support context, this might be less a problem. Since physicians are usually trained in EBM, we can require that physicians always formulate their information need in the EBM query frame. In fact, the rationale for advocating EBM is that by formulating the structured queries, physicians will be able to reflect more on their information need, thereby improving the quality of the clinical decision processes and patient care. Still, if one really wants a fully automated system that can do free-text search with this knowledge-based framework, then obviously a sophisticated information extraction module needs to be applied to the free-text query first in order to convert it into an EBM-style structured query.

7.2 Semantic-Based Query/Document Preprocessing

Many of the systems that participated in the TREC 2011 and TREC 2012 Medical Records track that achieved good performance utilized semantic mapping (e.g. MetaMap) or computer-based medical coding tools. Many of these systems convert documents and queries into "bags of concepts" by running them through semantic mapping tools and then perform concept-based indexing and retrieval. This approach was more feasible for the 2011 and 2012 tasks because the sizes of the corpora were much smaller. We estimated that running MetaMap on all 733K documents in the TREC 2014 CDS corpus would take weeks to months so we investigated other options.

We believe that our query expansion approach best approximates this "bag of concepts" strategy given our constraints. Indeed, in our query expansion step, we group all string variants corresponding to a concept as synonyms. Any string variant in the document matches this synonym equally, and so this approach is almost equivalent to that which first converts the terms in the documents into concepts and then matches them with the concepts in the query.

In Atigeo's participation in the TREC 2012 Medical Records track (Tinsley et al. 2013) a task-specific method of semantic-based document preprocessing was explored. The system first extracts the ICD-9 codes from the EMRs and then enriches the medical records with the text descriptions of the ICD-9 codes and their parent codes before indexing them. We conducted similar explorations on the MEDLINE corpus, in which each document is indexed with human-assigned MeSH terms. Some experiments showed that enriching the documents with the text descriptions – or "scope notes" – of the MeSH terms could lead to significant performance gain in document retrieval. However, as the MeSH terms are not included in the PubMed Central corpus, this approach would not work for the TREC 2014 CDS task. An alternative approach is to add the text descriptions of the terms identified by MetaMap to the queries, but our experiments suggested this would introduce too much noise and harm the result.

Another common query expansion technique is pseudo-relevance feedback. We have found the effect of pseudo-relevance feedback to be very sensitive to the numerical values of its parameters. Since we did not have many topics that we could tune our system on, we decided not to pursue this approach. There is also a query expansion approach to be discussed in Section 7.4 that combines semantic information and pseudo-relevance feedback, but sadly, that approach also only works for the MEDLINE corpus and not for the PubMed Central Corpus.

7.3 Advanced NLP Techniques for Medical Records Search

There have been several follow-up studies of the Medical Records track beyond the TREC 2011 and TREC 2012 submissions. With the labeled data from 2011 and 2012 available, and free from the time limitations imposed on TREC participants, these studies were able to pursue the applications of advanced (usually statistical) NLP techniques to the Medical Records track. It can be expected that similar studies will be spurred up for the CDS track.

Limsopatham et al. (2013a) addressed the issue that there are richer relationships between the med-

ical concepts beyond the hierarchical order specified by thesauri such as MeSH. For example, we might have some drug that is primarily used for treatment of certain diseases, but the drug and the disease concepts in the thesauri are usually unrelated. The authors developed an inference framework where this kind of relationship can be inferred statistically from the EMR corpus. Although the authors did not report significant improvement on the cohort selection task, this approach might be effective for the CDS task. We might, for instance, add in the concepts of drugs that are derived statistically in this framework from the diseases present in the query, if the query is of the `treatment` type.

In Limsopatham et al. (2013b), the authors evaluated a strategy that aggregates the results of the term-based and concept-based retrieval models. The aggregated score for each document is a weighted sum of the scores from both models, where the weights are estimated by supervised learning. An interesting feature of the aggregation strategy is that the weights are different for each query. Conceptually, this means that the aggregation function tries to infer the relative importance of the words and of the concepts for a given query. Again, although it did not yield significant improvement on the cohort selection task, it might be effective for the CDS task.

A more recent study of the Medical Records track was Wang et al. (2014), in which the authors propose an axiomatic method to regularize the weights of the concepts in the "bag of concept" representations, and experiments showed that this weighting scheme did improve the retrieval results.

7.4 Studies on Medical Document Retrieval

Medical document retrieval has been an active research area for decades. This section surveys some studies on general-purpose document retrieval, focusing on efforts that have been made to improve search results beyond basic term-based retrieval. This line of work, however, has mostly been done with respect to the MEDLINE corpus, and although they provide interesting insight to the problem, many of the methods may not be applicable to other corpora such as PubMed Central.

A series of experiments by Haynes et al. (1994, 2004, 2005) investigated the best strategy for retrieving documents pertaining to different clinical tasks (e.g. treatment or diagnosis), focusing on creating search rules that produce results with either the highest sensitivity or specificity. This effort has evolved into the PubMed Clinical Queries tool[10]. The rules function as filters on the search results and thus can be combined with other boolean retrieval or ranked retrieval queries that suit the user's information need. However, some of the rules devised by Haynes et al. include restriction on the MeSH terms assigned to the documents and thus apply only to searching the MEDLINE corpus. Even for those rules that only involve constraints on article titles and abstracts, the evaluation in the papers were only done based on the results of MEDLINE search. The performance of these filtering rules for searching other medical databases merits further critical evaluation.

Srinivasan (1996a) explored different query expansion methods for querying MEDLINE and discovered that adding MeSH terms in a pseudo-relevance fashion yields a significant improvement. This query expansion is done in two steps. First, a set of documents is first retrieved with the original query. Then the MeSH terms assigned to the top-ranked documents retrieved in the first step are added, and the final documents are retrieved using the expanded query. This method only works for MEDLINE, but a similar idea can be applied when querying other corpora. For example, we can add the terms found in the keywords fields of the top-ranked documents retrieved using the original query. However, in the PubMed Central corpus, we noticed that many of the documents do not have the keyword field. Therefore, we did not investigate how well this strategy might benefit performance on the CDS task.

Srinivasan (1996b) conducted experiments on the MEDLINE corpus with more indexing and retrieval strategies, among which the best result was achieved by expanding the query with MeSH terms in the same fashion as described in Srinivasan (1996a), then querying a free-text index and another MeSH index, and finally aggregating the results using a weighted sum of the scores.

[10] http://www.ncbi.nlm.nih.gov/pubmed/clinical/

Acknowledgments

The authors are grateful for the contributions of many of our colleagues at Atigeo in supporting our work on TREC 2014, especially the assistance of Bryan Tinsley, a veteran of our TREC 2012 team.

References

Aronson, A. R. and Rindflesch, T. C. 1997. Query Expansion Using the UMLS Metathesaurus. Proc. AMIA Annu. Fall Symp. 1997, 485–489.

Aronson, A. R. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proc. AMIA Symp. 2001, 17–21.

Demner-Fushman, D. and Lin, J. 2007. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. Computational Linguistics 33(1):63–103.

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Rance, B., Lang, F.-M., Ide, N., Apostolova, E., and Aronson, A. R. 2012. A Knowledge-Based Approach to Medical Records Retrieval. In Proceedings of the 20th Text REtrieval Conference (TREC 2011).

Demner-Fushman, D., Abhyankar, S., Jimeno-Yepes, A., Loane, R., Lang, F., Mork, J. G., Ide, N., and Aronson, A. R. 2013. NLM at TREC 2012 Medical Records Track. In Proceedings of the 21st Text REtrieval Conference (TREC 2012).

Haynes, R. B. Wilczynski, N. L., McKibbon, K. A., Walker, C. J., and Sinclair, J. C. 1994. Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE. J. Am. Med. Inform. Assoc. 1(6):447–458.

Haynes, R. B. and Wilczynski, N. L. 2004. Optimal Search Strategies for Retrieving Scientifically Strong Studies of Diagnosis from MEDLINE: Analytical Survey. BMJ 328(7447):1040.

Haynes, R. B., McKibbon, K. A., Wilczynski, N. L., Walter, S. D., and Werre, S. R. 2005. Optimal Search Strategies for Retrieving Scientifically Strong Studies of Treatment from MEDLINE: Analytical Survey. BMJ 330(7501):1179.

Krovetz, R. 1993. Viewing Morphology as an Inference Process. Proc. 16th Annu. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'93), 191–202.

Limsopatham, N., Macdonald, C., Ounis, I., McDonald, G., and Bouamrane, M.-M. 2012. University of Glasgow at Medical Records Track 2011: Experiments with Terrier. In Proceedings of the 20th Text REtrieval Conference (TREC 2011).

Limsopatham, N., Macdonald, C., and Ounis, I. 2013a. Inferring Conceptual Relationships to Improve Medical Records Search. Proc. 10th Conf. on Open Research Areas in Information Retrieval (OAIR 2013), 1–8.

Limsopatham, N., Macdonald, C., and Ounis, I. 2013b. Learning to Combine Representation for Medical Records Search. Proc. 36th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'13), 833–836.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. 1995. The Well-Built Clinical Question: A Key to Evidence-Based Decisions. ACP J. Club 123(3):A12–13.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., and Richardson, W. S. 1996. Evidence Based Medicine: What It Is and What It Isn't. BMJ 312(7023):71–72.

Srinivasan, P. 1996a. Query Expansion and MEDLINE. Inform. Process. Manage. 32(4):431–443.

Srinivasan, P. 1996b. Optimal Document-Indexing Vocabulary for MEDLINE. Inform. Process. Manage. 32(5):503–514.

Tinsley, B., Thomas, A., McCarthy, J. F., and Lazarus, M. 2013. Atigeo at TREC 2012 Medical Records Track: ICD-9 Code Description Injection to Enhance Electronic Medical Record Search Accuracy. In Proceedings of the 21st Text REtrieval Conference (TREC 2012).

Wang, Y., Liu, X., and Fang, H. 2014. A Study of Concept-Based Weighting Regularization for Medical Records Search. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 603–612.